Heriot Watt University

Data Mining and Machine Learning

Coursework 1

By

Vytautas Tumas

31 October 2015

1. Field distribution histogram

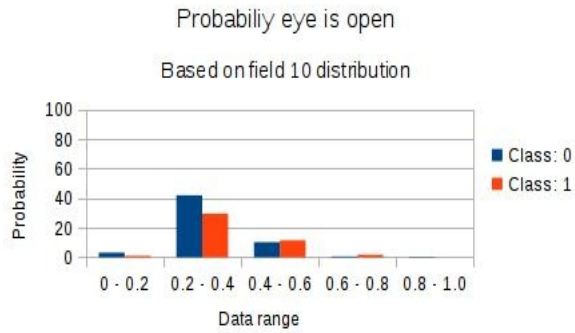

Illustration 1: Field 10

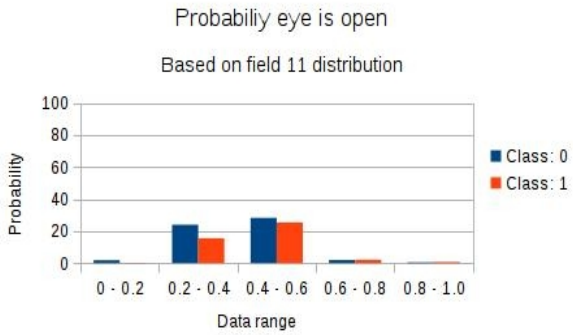

Illustration 2: Field 11

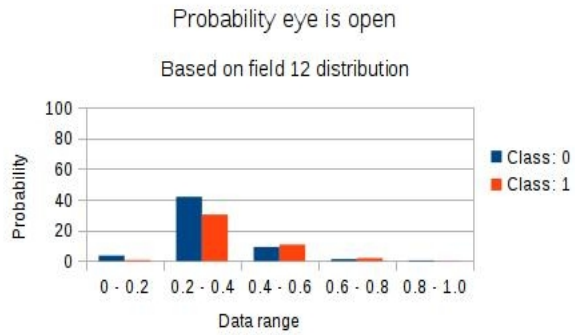

Illustration 3: Field 12

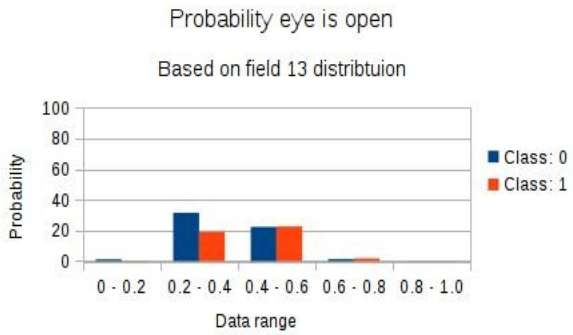

Illustration 4: Field 13

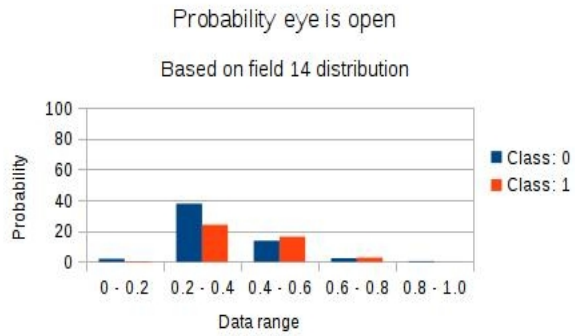

Illustration 5: Field 14

The histograms for EEG eye state dataset depict the state of the eye (open or closed) given the value distribution of the field. We can observe that in field 10, 12, 13, 14 majority of the values are in the range between 0.2 and 0.4. Furthermore, in that range, the eye is closed more frequently. In Illustration 1 we can see that 71% of the values are between 0.2 and 0.4 and 41% of the times the eye is closed and opened 30% of the time. 22% of the values are between 0.4 and 0.6, here 12% of the values indicate the eye is open and 10% indicate it's closed.
In Illustration 2 we can see that 53% of the values are in the range 0.4 - 0.6 and 39% of the values are in range 0.2 – 0.4.

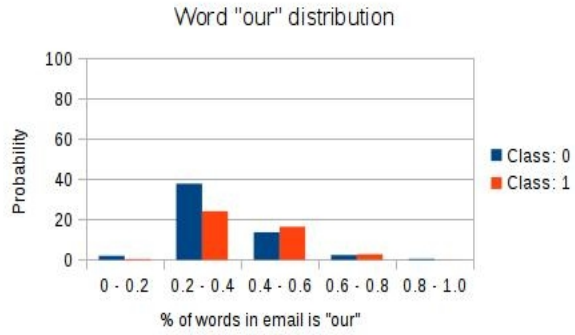

Illustration 6

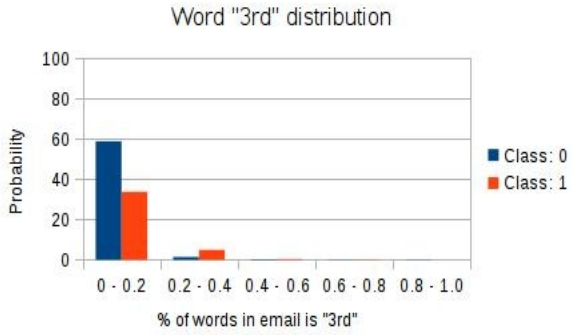

Illustration 7

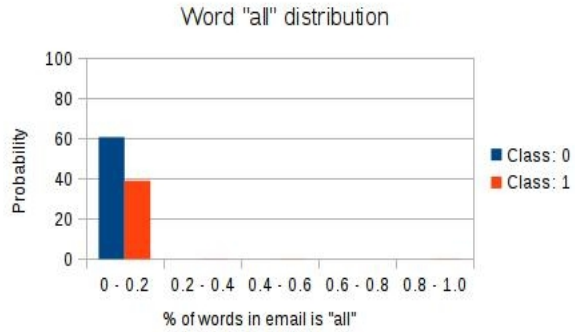

Illustration 8

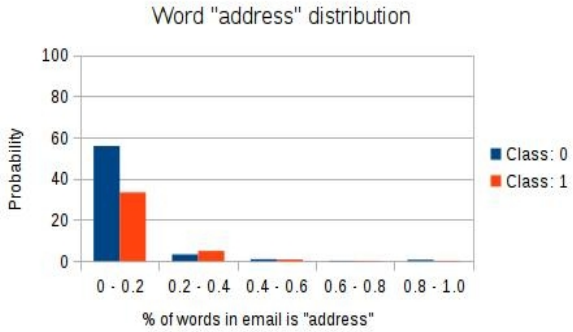

Illustration 9

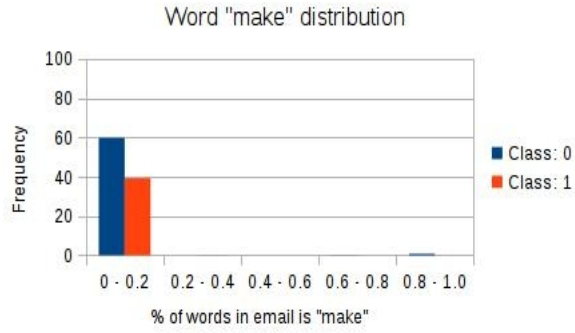

Illustration 10

The histograms for spambase dataset depict the number of emails marked as spam or non spam given how often the given word appears in the text. Illustrations 7 to 10 show that these words rarely appear in an email and when the word appears in the mail, the chances of it being spam are lower than it being a non spam email. In the Illustration 6 we can see that when word “our” distribution is between 0.4 and 0.6, 16% of the emails have been marked as spam and around 13% of the emails have been marked as non spam.

2.1 EEG Eye State histogram with removed highest bound

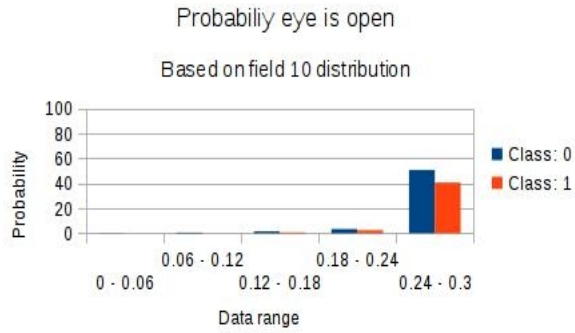

Illustration 11

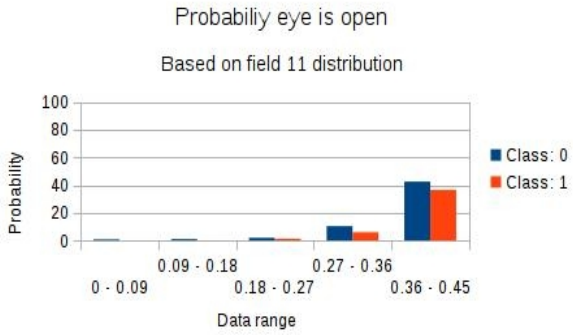

Illustration 12

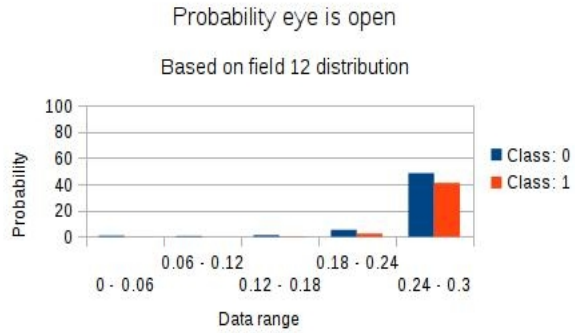

Illustration 13

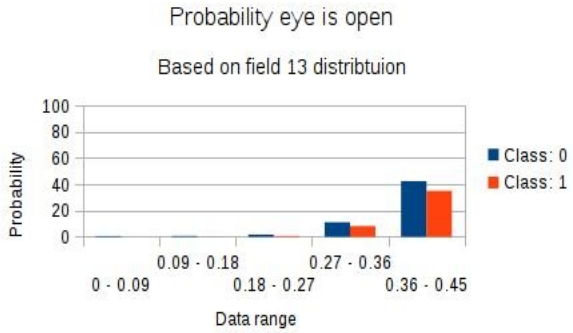

Illustration 14

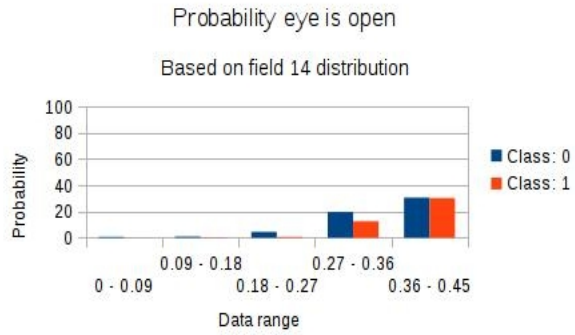

Illustration 15

For all of the histograms depicted above, the bin range was calculated after discarding the highest value in the field. As the result, we can see a more accurate value distribution for each field. Most of the values in Field 10 and 12 are in the range between 0.24 and 0.3, and around 50% of the records state the eye is closed and it's open in 40% of the records.

Values in fields 11, 13 and 14 mainly fall in to range between 0.36 and 0.45. With the exception of field 14, the eye is closed in around 40% of the cases and opened in 30-35% of the cases. In field 14 the eye is opened and closed in the same % of records.

2. Nearest Neighbour algorithm accuracy

```
tapanito@tapanito:~/university/dmmi/cwl/code/eeg$ python nn.py min_maxed
Working on: min_maxed
Accuracy of nearest neigbour is: 89.13%
```

Illustration 16: EEG Eye State Nearest Neighbour accuracy

```
Accuracy of nearest neigbour is: 61.1%. Field 13 Field 12
Highest accuracy is: 62.11% for Field 0 and Field 6
Average accuracy is: 56.33%
tapanito@tapanito:~/university/dmmi/cwl/code/eeg$
```

Illustration 17: Reduce EEG Eye State Nearest Neighbour accuracy

When running nearest neighbour algorithm on the full dataset we get higher accuracy than when running the same algorithm on column reduced dataset. This is because the distance is calculated using Euclidean distance, thus the accuracy is highly affected by the number of fields used for the algorithm. For EEG Eye State dataset the algorithm correctly guesses the class 89.1% of the time.
I could not determine which two fields in the dataset would provide the highest accuracy thus I ran an algorithm to try all possible combinations. I have determined that Field 1 and Field 7 have the highest accuracy of 62.11%. The average accuracy of the algorithm is 56.33%. Because of the nature of the algorithm used, the accuracy could be improved by using more fields in the algorithm.

```
tapanito@tapanito:~/university/dmmi/cwl/code/spambase$ python nn.py min_maxed
Working on: min_maxed
Accuracy of nearest neigbour is: 72.7%
```

Illustration 18: Spambase Nearest Neighbour accuracy

```
Accuracy of nearest neigbour is: 68.23%. Field 56 Field 55
Highest accuracy is: 72.85% for Field 51 and Field 54
Average accuracy is: 48.01%
tapanito@tapanito:~/university/dmmi/cwl/code/spambase$
```

Illustration 19: Reduced Spambase Nearest Neighbour accuracy

To reduce the running time of the nearest neighbour algorithm I have reduced the spambase dataset to 1149 records. The Nearest Neighbour algorithm correctly guesses the class 72.7% of the time.
Similarly to EEG Eye State dataset, I ran an the Nearest Neighbour for all possible combinations for two fields. I have determined that when using Field 52 and 55 the accuracy of the algorithm is 72.85%, which is 0.15% higher than the accuracy of the algorithm on the full data. The average accuracy is 48.01%.

3. Algorithms and Languages used

All the algorithms were implemented in Python. For each dataset the result of record reduction was piped to normalization script, then the result was passed to nearest neighbour algorithm and histogram generation. The histograms were made by exporting the bins and the distribution in csv format, which was then imported in Excel to make the graphs.